HW7 Group 1, Austin Halvorsen

Pedram Jahangiry

Nov 5 2020

# Problems

## Question 1

### (i)

If you have heteroskedasticity, you violate assumption MLR5 and cannot rely on the t stat

### (ii)

If you have a sample correlation coefficient of .95 between two independent variables in the model, you are not in violation of the assumptions and you can use the t stat.

### (iii)

If you omit an important explanatory variable, you introduce bias into your model, which would violate MLR.4, and you would not be able to rely on the t test for your model.

## Question 2

### (i)

The null hypothesis would be . The alternative hypothesis would be

### (ii)

I would expect to have a positive sign, since having a larger population would result in more demand for rental housing.

I would expect to also have a positive sign, since higher income would likely mean that the cost of living is high as well, meaning housing prices and rental rates would be higher.

### (iii)

Because this is a log-log model, the interpretation of this model should be, \*"A 10% increase in population would result an increase of rent of 0.66%.

### (iv)

With a degree of freedom of 60 (64-4), at the 1% level, we need to calculate the t value as

This would give an estimated t stat of

qt(0.995, 60)

## [1] 2.660283

We would reject the null hypothesis at the 1% significance level.

## Question 3

### (i)

We fail to reject the null hypothesis where and we also fail to reject the null hypothesis where . We calculate this be evaluating the t stat for each parameter:

and

### (ii)

We need to calculate the F statistic where:

So we are going to reject the null that they are jointly significant.

### (iii)

To test the joint hypothesis that $H\_0:\beta\_2=0, \beta\_3=0, and \space \beta\_4=0$ we need to test

So, with this output, we fail to reject the null.

### (iv)

You cannot rely on the F statistic from part iii because we have violated assumption MLR.5.

## Question 4

### (i)

We can apply the property of variance:

### (ii)

So to identify the t statistic, we get:

### (iii)

If then , therefore we can write our population regression model as:

We would estimate y by regressing y on and . This will allow us to calculate the stand error and coefficient of

## Question 5

### (i)

With our t stat close to 1: (), we will fail to reject that this variable is irrelevant.

### (ii)

Yes, the effect is significant. The t value is greater than 2, and a critical value of and this would make sense that companies with larger market values pay more.

### (iii)

The t value for ceoten is and the t value for comten so they are both significant to our model.

### (iv)

This could be the result of someone who has stayed with the company for a long time and may not have demands on pay and does the job more out of passion for the company rather than an outsider who may demand large bonus pay to run the company.

# Computer Exercises

## Question 6

### (i)

d1 <- hprice1  
mrm1 <- lm(lprice~sqrft+bdrms, d1)  
stargazer(mrm1, type = 'text', digits = 8)

===============================================  
 Dependent variable:   
 ---------------------------  
 lprice   
-----------------------------------------------  
sqrft 0.00037945\*\*\*   
 (0.00004321)   
   
bdrms 0.02888444   
 (0.02964326)   
   
Constant 4.76602700\*\*\*   
 (0.09704447)   
   
-----------------------------------------------  
Observations 88   
R2 0.58829480   
Adjusted R2 0.57860770   
Residual Std. Error 0.19706340 (df = 85)   
F Statistic 60.72921000\*\*\* (df = 2; 85)  
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Using the data from HPRICE1, we get the following model:

with an and .

Using our model we get the following estimation:

theta <- 150\*mrm1$coefficients[2]+mrm1$coefficients[3]  
theta

sqrft   
0.08580134

This means that when we add a 150 sqft bedroom, then we can expect the home price to increase by about 8.6%

### (ii)

We can rewrite . In our model, we can substitute to have the following:

### (iii)

se1 <- lm(lprice~I(sqrft-150\*bdrms)+bdrms, d1)  
stargazer(se1, type = 'text')

==================================================  
 Dependent variable:   
 ---------------------------  
 lprice   
--------------------------------------------------  
I(sqrft - 150 \* bdrms) 0.0004\*\*\*   
 (0.00004)   
   
bdrms 0.086\*\*\*   
 (0.027)   
   
Constant 4.766\*\*\*   
 (0.097)   
   
--------------------------------------------------  
Observations 88   
R2 0.588   
Adjusted R2 0.579   
Residual Std. Error 0.197 (df = 85)   
F Statistic 60.729\*\*\* (df = 2; 85)   
==================================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

pander(confint(se1, level = 0.95))

|  |  |  |
| --- | --- | --- |
|  | 2.5 % | 97.5 % |
| **(Intercept)** | 4.573 | 4.959 |
| **I(sqrft - 150 \* bdrms)** | 0.0002935 | 0.0004654 |
| **bdrms** | 0.03258 | 0.139 |

## Question 7

### (i)

d2 <- k401ksubs  
d2 %>%   
 filter(fsize==1) %>%   
 count()

n  
1 2017

There are 2017 single person households in the dataset.

### (ii)

d2a <- d2 %>%   
 filter(fsize==1)  
mrm2 <- lm(nettfa~inc+age, d2a)  
stargazer(mrm2, type='text')

===============================================  
 Dependent variable:   
 ---------------------------  
 nettfa   
-----------------------------------------------  
inc 0.799\*\*\*   
 (0.060)   
   
age 0.843\*\*\*   
 (0.092)   
   
Constant -43.040\*\*\*   
 (4.080)   
   
-----------------------------------------------  
Observations 2,017   
R2 0.119   
Adjusted R2 0.118   
Residual Std. Error 44.683 (df = 2014)   
F Statistic 136.465\*\*\* (df = 2; 2014)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Using only single family households, we get the model:

With an and . This means that for every $1 thousand increase in income means an estimated $799 increase in net wealth. For we interpret that for every year increase, we estimate a $843 net wealth increase. These numbers make sense, and is not too surprising.

### (iii)

The intercept, , means that at age 0 and income = 0, their net wealth is $-43.04. In other words, you are born in debt.

### (iv)

The t-statistic is Against the one-sided alternative , the p-value is . So we can reject the null at the 5% significance level, but not the 1% significance level.

### (v)

srm <- lm(nettfa~inc, d2a)  
stargazer(srm, mrm2, type = 'text')

=======================================================================  
 Dependent variable:   
 ---------------------------------------------------  
 nettfa   
 (1) (2)   
-----------------------------------------------------------------------  
inc 0.821\*\*\* 0.799\*\*\*   
 (0.061) (0.060)   
   
age 0.843\*\*\*   
 (0.092)   
   
Constant -10.571\*\*\* -43.040\*\*\*   
 (2.061) (4.080)   
   
-----------------------------------------------------------------------  
Observations 2,017 2,017   
R2 0.083 0.119   
Adjusted R2 0.082 0.118   
Residual Std. Error 45.592 (df = 2015) 44.683 (df = 2014)   
F Statistic 181.599\*\*\* (df = 1; 2015) 136.465\*\*\* (df = 2; 2014)  
=======================================================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

When calculating the regression using just inc, the estimated coefficient is larger than it is the first model. It is going to be larger because the correlation is small between age and inc. So when we remove the age, we introduce a small postivie bias into the model which gives us a larger estimated coefficient.

## Question 8

### (i)

d3 <- discrim  
mrm3 <- lm(lpsoda~prpblck+lincome+prppov, d3)  
stargazer(mrm3, type='text')

===============================================  
 Dependent variable:   
 ---------------------------  
 lpsoda   
-----------------------------------------------  
prpblck 0.073\*\*   
 (0.031)   
   
lincome 0.137\*\*\*   
 (0.027)   
   
prppov 0.380\*\*\*   
 (0.133)   
   
Constant -1.463\*\*\*   
 (0.294)   
   
-----------------------------------------------  
Observations 401   
R2 0.087   
Adjusted R2 0.080   
Residual Std. Error 0.081 (df = 397)   
F Statistic 12.604\*\*\* (df = 3; 397)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

We get the following model:

With an and . AT the 5% level, is statistically significant, but it is not significant at the 1% level.

### (ii)

pander(cor.test(d3$lincome,d3$prppov, method = "pearson", use="complete.obs"))

Pearson’s product-moment correlation: d3$lincome and d3$prppov

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test statistic | df | P value | Alternative hypothesis | cor |
| -31.04 | 407 | 2.349e-109 \* \* \* | two.sided | -0.8385 |

summary(mrm3)

Call:  
lm(formula = lpsoda ~ prpblck + lincome + prppov, data = d3)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.32218 -0.04648 0.00651 0.04272 0.35622   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1.46333 0.29371 -4.982 9.4e-07 \*\*\*  
prpblck 0.07281 0.03068 2.373 0.0181 \*   
lincome 0.13696 0.02676 5.119 4.8e-07 \*\*\*  
prppov 0.38036 0.13279 2.864 0.0044 \*\*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.08137 on 397 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.08696, Adjusted R-squared: 0.08006   
F-statistic: 12.6 on 3 and 397 DF, p-value: 6.917e-08

The correlation between log(income) and prppov is **-0.8385**.

The p-value for log(income) is 0.0045 and the p-value for prppov is 0.00001. These are both significant at the 1% level.

### (iii)

mrm4 <- lm(lpsoda~prpblck+lincome+prppov+lhseval, d3)  
stargazer(mrm4, type='text')

===============================================  
 Dependent variable:   
 ---------------------------  
 lpsoda   
-----------------------------------------------  
prpblck 0.098\*\*\*   
 (0.029)   
   
lincome -0.053   
 (0.038)   
   
prppov 0.052   
 (0.134)   
   
lhseval 0.121\*\*\*   
 (0.018)   
   
Constant -0.842\*\*\*   
 (0.292)   
   
-----------------------------------------------  
Observations 401   
R2 0.184   
Adjusted R2 0.176   
Residual Std. Error 0.077 (df = 396)   
F Statistic 22.313\*\*\* (df = 4; 396)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

summary(mrm4)

Call:  
lm(formula = lpsoda ~ prpblck + lincome + prppov + lhseval, data = d3)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.30652 -0.04380 0.00701 0.04332 0.35272   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.84151 0.29243 -2.878 0.004224 \*\*   
prpblck 0.09755 0.02926 3.334 0.000937 \*\*\*  
lincome -0.05299 0.03753 -1.412 0.158707   
prppov 0.05212 0.13450 0.388 0.698570   
lhseval 0.12131 0.01768 6.860 2.67e-11 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.07702 on 396 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.1839, Adjusted R-squared: 0.1757   
F-statistic: 22.31 on 4 and 396 DF, p-value: < 2.2e-16

Adding the variable **log(hseval)** we get the following model:

Using this model, we can interpret the coefficient to mean that a 1% higher median housing value in a zip code with increase the predicted price of a medium soda by 0.121%.

For the two-sided p-value at we have a low p-value.